DOCUMENT RESUME

ED 266 690 FL 015 546

AUTHOR Shohamy, Elana; And Others
TITLE From Testing Research to Educational Policy: Introducing a New Comprehensive Test of Oral Proficiency.
PUB DATE 85
NOTE 13p.; In: Language Testing in School. AFinLA Yearbook 1985. No. 41 (see FL 015 537). Document is in small print.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Foreign Countries; Group Discussion; Interviews; *Language Proficiency; *Language Styles; *Language Tests; Role Playing; Secondary Education; Second Language Instruction; *Speech Communication; Student Attitudes; Test Reliability
IDENTIFIERS *Israel

ABSTRACT

A study was designed to develop a number of tests of oral proficiency and to compare those tests with the existing, highly subjective testing method used at the end of secondary school in Israel. The study used four experimental tests: the oral interview, role playing, reporting task, and group discussion. The experimental tests were analyzed for the extent to which they discriminate among levels of oral proficiency, interrater reliability, and student attitudes. Comparison with the traditional test was based on mean scores, standard deviation, distribution of scores, and correlations. In addition, the relationships among the four experimental tests regarding their representation of different speech styles were examined. It was found that the experimental tests were appropriately discriminatory as a group, that they had content validity, and that student attitudes toward them were favorable. They were found to be preferable to the existing oral test for providing meaningful information; some evidence that they address different constructs of oral proficiency was also found. It is recommended that new tests be introduced into the national educational system. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Tomala

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED266690

FROM TESTING RESEARCH TO EDUCATIONAL POLICY: INTRODUCING A NEW COMPREHENSIVE TEST OF ORAL PROFICIENCY

Elana Shohamy (Tel-Aviv University)
Thea Reves (Bar Ilan University)
Yael Bejarano (Everyman's University)

1 Introduction

In Israel oral proficiency has been tested within the framework of the high-school leaving examination for twenty years. An oral test is administered by the Ministry of Education to all students at the end of the 12th grade. The common procedure is an informal, unstructured interview, conducted by a high school teacher who, by virtue of his experience in teaching, is called upon to administer the test. The topics of the conversation between the tester and the test-taker vary freely from everyday matters to more sophisticated subjects, without any advance planning. The scores assigned to the test-taker as a result of this 'test' are totally impressionistic, based on the tester's subjective judgement.

A number of deficiencies can be found in this procedure: a) It represents a single aspect of oral performance, and it is therefore questionable whether it is a valid indication of students' overall oral proficiency. b) Since rater reliability is not measured, it is questionable whether the score obtained by the student is his/her true score, especially since the testers are not trained in either interviewing techniques or the rating of oral proficiency. c) The test has low variance and a narrow distribution of scores: literally nobody fails the test. It is clear that the test is not sensitive enough to tap the existing individual differences in oral proficiency.

Searching for an alternative to the existing system constitutes a problem, since among the large number of tests available (Madsen and Jones, 1981), hardly any have been sufficiently researched to allow their implementation on a nation-wide scale. The only oral test that has been researched extensively is the FSI (Foreign Service Institute) Oral Interview (Bachman and Palmer, 1981, Shohamy 1982, Hinofotis 1976, Clifford 1977). However, one of the major shortcomings of that test is that it does not encompass a wide enough variety of speech styles (Shohamy 1983); it is limited to questions asked by the tester and answers supplied by the test-taker. It is obvious that such a test cannot be a valid measure for assessing all aspects of oral proficiency. Recognizing this shortcoming, a number of U.S. government agencies have recently incorporated additional speech styles into the FSI Oral Interview procedure, such as reporting, conversing, and role playing (Lowe 1982). To date, however, there are no research data available on this newly revised test.

FL 015546

BEST COPY AVAILABLE

It is obvious that adding other speech styles to the Oral Interview can be justified from a linguistic point of view, since various speech styles utilize different linguistic and socio-linguistic features in oral discourse.

2 The purpose of the study

The aim of the research study presented in this paper is as follows:
1) To develop a number of experimental oral proficiency tests. The tests are intended to a) represent a broader range of oral speech styles which would presumably reflect the type of communicative language that students are likely to need in authentic situations, b) discriminate well among various levels of students' oral proficiency, c) have higher inter-rater reliability. These tests should also evoke positive attitudes on the part of the test-takers.

2) To examine the relationship between the scores obtained on the experimental tests and the scores on the existing oral proficiency test. It was assumed that if the psychometric quality of the experimental tests proves to be better, they may be considered a preferable alternative to the existing test.

3) To examine the extent to which the experimental tests are related to one another. In other words, whether any one of the tests can be a valid measure of overall oral proficiency, or whether there is a justification for the testing of the various speech styles by separate oral tests.

Four experimental tests were designed: Oral Interview, Role Play, Reporting Task and Group Discussion. The experimental tests were systematized in three ways: a) the tasks were constructed so as to discriminate among levels of performance; b) testers were trained to conduct and rate the tests, and c) scores were based on a systematic rating scale. These experimental tests were then compared to the existing oral test used in the national school leaving examination, for their psychometric properties.

This paper will report research results obtained on the development of the experimental tests, their relationship to the existing conventional test, and the inter-relationship among the experimental tests. These findings will provide the basis for recommendations regarding policy decisions.

Specifically, the following questions were asked:
1. Regarding the experimental tests: To what extent do they discriminate among various levels of oral proficiency and what is their inter-rater reliability? What are students' attitudes towards the tests?
2. Regarding the comparison between the experimental oral tests on the one hand, and the existing oral test on the other: Is there a difference between the means, standard deviation, distribution of scores, and correlation of the two types of tests?

3. Regarding the relationship among the four experimental tests: To what extent do students' scores on the different types of oral tests represent different speech styles related to one another and does the degree of relationship, by itself, suggest the need for different tests?

3 The design

The sample of the study consisted of one hundred and three twelfth grade students of four classes in a comprehensive high school north of Tel Aviv. These classes were randomly chosen out of seven classes in that school. All the students undertook all the experimental oral tests - the Oral Interview (OI), the Role Play (RPL), the Reporting Task (REP), and the Group Discussion (GD). The tests were administered on the campus of the Tel Aviv University; each test was administered independently and lasted for about ten minutes. In order to minimize the learning effect possibly created by the order of the tests, groups of students took the tests in various orders, so that total rotation was ensured. The OI, the RPL, and the REP were carried out on a one-to-one basis, while the GD was held in groups of four. Different testers were assigned to administer the tests; they were all experienced EFL teachers and were trained in administering and rating the tests. The rating of the oral performance was done on the spot, using a rating scale which was adapted from several existing rating scales, mostly from the DLI (Defense Language Institute).

It rated oral proficiency on a scale ranging from 4 to 10*, 10 being equivalent to native speaker's performance. The OI, RPL, and REP tests were audio-taped, and the GD was video-taped. This was done in order to provide the data for computing rater reliability, that is, comparing the scores assigned on the spot with scores given at a later date by independent, lay raters.

On completion of each of the four tests the students filled out a questionnaire which assessed their attitudes toward the four experimental tests.

Two weeks after the administration of the four experimental tests 77 of the 103 students were tested by the existing conventional test in their schools. It is on these 77 cases that the comparison between the experimental tests and the existing test was done.

3.1 The experimental tests

Four tests were used as the experimental oral test. The rationale for the selection of the experimental tests in this study was to tap a variety of speech styles which are likely to be needed in various oral communicative situations. It is realized that although these situations do not embrace the whole scope of oral

*The scale of 4 to 10 is the conventional scale regularly used in the Israeli School System.

performance, they provide a representative sample of useful oral interactions.

a) The Oral Interview (OI):
The rationale underlying this test was to guide the test-taker into a dialogue with the tester, which elicits answers to questions asked by the tester on different topics at different levels of proficiency. The test followed the model of the Foreign Service Institute Oral Interview (Lowe 1982). It was administered by trained testers and took about 10 minutes; in included the following four phases: 1) warm up; in this phase the test-taker was put at ease and the tester derived a preliminary indication of the test-taker's level of proficiency in speech and understanding. 2) level-check: during this phase the tester checked the functions and content which the test-taker could perform most accurately. 3) probing: in this phase the tester assessed the highest level at which the candidate could function accurately. 4) wind-up: in this phase the test-taker was returned to the level at which he could function most comfortably and thus was left with a feeling of accomplishment. The scoring of the test-taker's performance on the oral interview was done on the basis of the rating scale used for all the other tests.

b) The Role Play (RPL):
The rationale behind this test was to stimulate the test-taker to produce spontaneous speech-behavior within the limits of a pseudo-authentic situation on the one hand, and to elicit specific speech acts prompted by strictly limited situations on the other. The test was composed of a situation in which the test-taker had to play a role with the tester playing another, both partners in a dialogue. The test-taker was given a card on which he found the description of a situation and his expected role in it. The card was written in the test-taker's mother tongue, Hebrew, so that interference of reading comprehension in L2 should be avoided. The tester then engaged in the simulated conversation derived from the situation described on the card. The test lasted for about ten minutes, and a score was assigned by an assessor who was not involved in the RPL, on the basis of the rating scale used for all the other tests.

c) The Reporting Task (REP):
The rationale underlying this test was to stimulate the test-taker into a monologue in the foreign language based on authentic input in the mother tongue, Hebrew. The student was given an article in Hebrew which he was asked to read silently, and report its general content in his own words, in English. He was specifically asked not to translate the text, but rather to report freely, referring back to the text only if necessary. The test lasted about 10 minutes and was scored on the basis of the rating scale used to rate oral proficiency for all the other experimental tests.

d) The Group Discussion(GD):
The rationale underlying this test was to stimulate the test-takers

into a spontaneous discussion of a controversial issue, in which they could express views about topical matters, debate and argue over them, defend their opinions and try to persuade the other participants to accept them. Four students were asked to discuss a topical subject or issue controversial enough to lend itself to a lively discussion. Members of the group picked a card randomly from among 20 cards on the table. The card provided information regarding the topic of the discussion they were about to conduct and some clues for the discussion, such as guiding questions with some relevant lexical items. They were given a few minutes to read the card and plan the procedure of their discussion among themselves in their mother tongue, Hebrew, before they were to start the actual discussion in English. Then they discussed the topic in English (Reves 1982). The tester listened to the discussion without participating in it or interfering with it, and scored the performance of each of the four participant test-takers on the basis of the rating scale used for all the other tests in this study.

4 Results

4.1 Regarding the experimental tests:

a) Distribution of Scores:
Table 1 displays the distribution of scores obtained by the students on the four experimental tests. It shows that the distribution of scores for the four tests ranged along the whole scale of 4 to 10. In general, the distribution of scores on the experimental tests was similar.

Table 1

DISTRIBUTION OF SCORES ON THE EXPERIMENTAL TESTS

	OI		RPL		REP		GD		AVERAGE	
Score	f	%	f	%	f	%	f	%	f	%
4	2	1.9	3	2.9	1	1.0	9	8.7	4	3.62
5	9	8.7	13	12.6	10	9.7	9	8.7	10	9.93
6	27	26.2	31	30.1	30	29.1	27	26.2	29	27.9
7	38	36.9	33	32.0	34	33.0	40	38.8	36	34.9
8	11	10.7	10	9.7	11	10.7	7	6.8	10	9.48
9	10	9.7	8	7.8	12	11.7	5	4.9	9	8.63
10	6	5.8	5	4.9	5	4.9	6	5.8	5	5.40

N = 103

b) Inter-rater Reliability:
The extent of agreement of different raters regarding the oral proficiency scores was measured by correlating scores assigned by the tester on the spot (at the completion of the test) with a

second rating done on a stratified random sample of 25 selected tapes of each of the tests. It was carried out by two independent, untrained raters using the same rating scale as used for the first rating. The correlations were .91 for the OI, .81 for the REP, .76 for the RPL and .73 for the GD test, as displayed in Table 2. It is clear from these analyses that the reliability of the tests was relatively high, although it varied from one test to the other; the GD had the lowest and the OI the highest. We would assume that by assuring proper training of testers the inter-rater reliability could be even further improved.

Table 2

INTER-RATER RELIABILITY OF THE FOUR ORAL TESTS*

TEST	r
OI	.91
RPL	.76
REP	.81
GD	.73

*Based on 25 samples of each test

c) Attitude towards the four oral tests:
Attitude towards the four oral tests was assessed using a questionnaire that students filled out upon completion of each test. In the questionnaire students were asked to mark their agreement with statements which described the test as being difficult, interesting, instructive, etc. on a scale of 1 to 4, 4 meaning "strongly agree". Table 3 presents these results indicating that students favored the OI and the RPL: $\bar{x} = 3.05$ and $\bar{x} = 3.03$, respectively. The attitude towards the other two tests was less favorable, although still positive ($\bar{x} = 2.59$ for the Reporting and $\bar{x} = 2.74$ for the Group Discussion).

Table 3

STUDENTS' ATTITUDES TOWARDS THE FOUR ORAL TESTS

TEST	$\bar{X}$	S.D.	Order of Priority
OI	3.05	.44	(1)
RPL	3.03	.36	(2)
REP	2.59	.512	(4)
GD	2.74	.43	(3)

n = 103

4.2 Regarding the comparison of the four experimental tests with the existing test:

a) Differences in means and standard deviations:

Table 4 displays the means and the standard deviations of the four experimental tests and of the existing test for the 77 students who took both types of tests. The means of the four experimental tests were quite similar: they ranged from $\bar{x}$ = 6.00 for the GD to $\bar{x}$ = 6.87 for the REP. In contrast to the experimental tests, the mean of the conventional test was substantially higher, $\bar{x}$ = 7.79, indicating the possibility that the scores may have been artificially inflated.

A substantial difference between the two types of tests was also found in the S.D. of the tests: while the S.D. of the experimental tests ranged from 1.32 to 1.93, the S.D. of the existing test was 1.03, indicating that the existing test is less sensitive to individual differences among students.

Table 4

MEANS AND STANDARD DEVIATIONS OF THE ORAL TESTS

Test	$\bar{X}$	S.D.
Oral Interview (OI)	6.49	1.39
Role Play (RLP)	6.17	1.51
Reporting (REP)	6.57	1.32
Group-Discussion (GD)	6.00	1.93
Existing Test	7.79	1.03

n = 77

A t-test analysis between the means of the existing test and each of the four tests indicated that these differences were statistically significantly different at the .001 level. Thus, a student may get a different score from the two types of tests.

b) Distribution of scores:
Table 5 displays the distribution of scores obtained by the students on the four experimental tests and on the existing test. It shows that while the distribution of scores for the four tests ranged along the whole scale of 4 to 10, the distribution on the existing test was very different: nobody received the scores of 4 and 5, while 13.55% of the students received these scores on the experimental tests. It is important to note that the scores of 4 and 5 are failing marks in the school system. It is worth noting what happened on the upper part of the scale. While about 56% of the students obtained scores of 8-9-10 on the existing test, only about 23% of the students received such scores on the experimental tests. The distribution of scores on the experimental tests did, in effect, resemble one another, but was still quite different on the existing test. These differences are even more noticeable in the polygraph displayed in Figure 1.

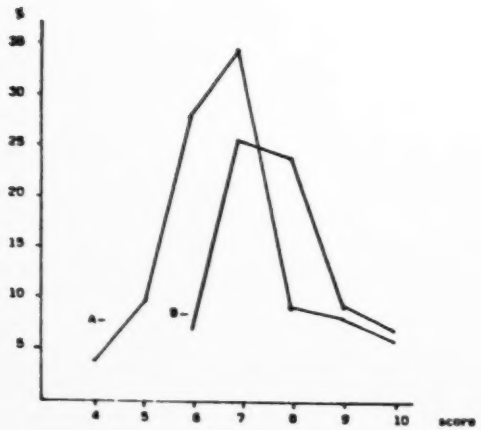
Table 5

COMPARISON OF DISTRIBUTION OF SCORES ON THE TWO TYPES OF ORAL TESTS

Score	Existing Test		Experimental Test (average of the four tests)	
	f	%	f	%
4	0	0	4	3.62
5	0	0	10	9.93
6	8	10.38	29	27.9
7	26	33.76	36	34.9
8	25	32.46	10	9.48
9	10	12.98	9	8.63
10	8	10.35	5	5.40
N = 77			N = 103	

Figure 1

DISTRIBUTION OF SCORES ON THE TWO TYPES OF ORAL TESTS



A : AVERAGE OF FOUR EXPERIMENTAL TESTS
B : EXISTING TEST

c) Relationship between the Experimental and the Existing Test: Table 6 displays the r and the r^2 of the experimental tests and the existing oral test. It shows that the correlations between the OI experimental test and the existing test, which are based on the same speech style, interviewing, was .63, thus explaining about 40% of the variance. The mutual variance of the existing test and the other tests which tap other speech styles of oral proficiency is even lower, r^2 = .23 with the RPL, r^2 = .29 with the REP, and r^2 = .20 with the GD. It is, therefore, clear that in spite of the fact that all the tests were tapping oral proficiency, it is probably the different speech styles or the different methods of testing that brought the mutual difference drastically down.

Table 6

r AND r^2 BETWEEN THE FOUR EXPERIMENTAL TESTS AND THE EXISTING TEST

Experimental / Existing test	r	r^2
OI	.63	.40
RPL	.48	.23
REP	.54	.29
GD	.45	.20

4.3 Examining the Four Experimental Tests:

a) Correlations among the four oral tests: Table 7 displays the r and the r^2 of the four experimental tests. The correlations ranged from .60 to .70, that is 36-49% mutual variance among the tests. The GD had the lowest mutual variance with the other tests, while the OI had the highest.

Table 7

r AND r^2 AMONG THE FOUR EXPERIMENTAL TESTS

TEST	RPL		REP		GD	
	r	r^2	r	r^2	r	r^2
OI	.69	.48	.70	.49	.60	.36
RPL			.67	.45	.64	.41
REP					.65	.42

5 Conclusions

Several conclusions can be drawn from these results. The first purpose was to develop and experiment with a number of oral proficiency tests which will represent a broad range of oral speech styles. These tests were expected to be accurate in terms of their psychometric properties and to evoke positive attitudes on the part of the test takers. The results obtained in this research indicate that the tests discriminated properly among the various levels of oral proficiency and had high inter-rater reliability (although it varied from one test to another). In addition, it can be claimed that by virtue of the fact that they tap a range of speech functions, the four tests, combined, have content validity. The test takers' attitudes to the tests proved to be favorable.

The second purpose of the research was to compare the experimental tests with the existing oral test. The existing test was found to have a restricted distribution of scores, limited variance, and the scores appeared to be inflated (practically nobody failed the test). In contrast the scores of the experimental tests covered the whole range of the rating scale, the variance was not restricted and the mean scores were more realistic. In other words, while meaningful information was obtained from the experimental tests, the existing test provided hardly any meaningful information. The correlations between the existing test and each one of the four experimental tests were not very high, even when the same speech style was tested. These findings allow us to conclude that, considering the above aspects, the experimental tests provide a preferable alternative to the existing oral test.

The third purpose of the research was to find out if and to what extent the relationship among the scores obtained on the four oral tests can provide evidence that the four speech styles are indeed different from a testing point of view. A low relationship could statistically justify the need for separate tests to tap overall oral proficiency. The mutual variance among the four experimental tests was found to be only moderate. This may provide evidence for the claim that the four tests indeed tap different constructs of oral proficiency.

6 Recommendations

The results of this research led us to recommend the introduction of a common measure of oral proficiency which consists of various speech styles. There is a strong educational advantage in expanding the number of speech functions in an oral test for high school students. It is most likely that the backwash effect created by a test encompassing multiple speech styles will give more emphasis to the instruction of these speech styles in the classroom. Clark (1980) calls for the development of such a test:

> "...the successful development of a comprehensive 'common measure' proficiency test would make available a highly

> refined and very powerful assessment instrument which could be used, in its own right, to evaluate communicative proficiency in situations for which the use of such an instrument would be considered both desirable and time - and cost - effective." (p. 25)

Regarding the structure of such a test which taps various language functions, there are two alternatives. One is a test composed of four isolated 'testing events', each conducted by a different tester (as for the experimental test in this study). Another alternative is a battery of tests, in which one tester tests the student in a number of functions in one 'testing event'. This latter version involves a separate rating for each speech function. An important question to investigate regarding these alternatives (isolated tests or a test-battery) is if the two types of oral test yield the same information for the test-taker.

7 From research to policy

The research reported in this paper demonstrates an example of the connection between research results and actual changes to be introduced into an educational system. Its main purpose was to evaluate an existing test and experiment with new tests, which, if found reliable and valid, would replace an existing testing procedure. The results of this research indeed indicate that there were good reasons to recommend the introduction of the new experimental tests.

The four-part oral proficiency test has indeed been accepted by the Ministry of Education to replace the existing oral exam. It will be introduced in 1986 as part of the nationwide Matriculation Examination taken by about 30,000 school-leaving high school students. It is an example of a situation in which the researcher can indeed be actively influential in creating a change in testing policy.

Bibliography

Bachman, Lyle F. and Adrian S. Palmer 1981. "The Construct Validation of the FSI Oral Interview". Language Learning, 31:67-86.

Clark, John L.D. 1980. "Toward a Common Measure of Speaking Proficiency". 15-26 in J.R. Frith (ed.). Measuring Spoken Language Proficiency. Washington, D.C.: Georgetown University Press.

Clifford, Ray T. 1977. Reliability and Validity of Language Aspects Contributing to Oral Proficiency of Prospective Teachers of German: In J.L.D. Clark (ed.). Direct Testing of Speaking Proficiency: Theory and Application. Princeton: Educational Testing Service.

Hinofotis, Frances B. 1976. An Investigation of the Concurrent Validity of Cloze Testing as a Measure of Overall Proficiency in English as a Second Language. Unpublished Ph.D. dissertation, Southern Illinois University.

Lowe, Pardee Jr. 1980. Manual for LS Interview Workshops. Washington, D.C., CIA Language School. 1982 (revised) (mimeo).

Madsen, Harold S. and Randall L. Jones 1981. "Classification of Oral Proficiency Tests", 15-30 in Adrian S. Palmer, Peter J.M. Groot and George A. Trosper (eds.). The Construct Validation of Tests of Communicative Competence. Washington, D.C. Teachers of English to Speakers of Other Languages.

Reves, Thea 1982. "The Group Oral Examination: A Field Experiment", World Language English, Vol. 1, no. 4:259-262.

Shohamy, Elana 1982. "Predicting Speaking Proficiency from Cloze Tests", Applied Linguistics, 3:161-171.

Shohamy, Elana 1983. "The Stability of Oral Proficiency Assessment on the Oral Interview Testing Procedure", Language Learning, 33:3.

Wilds, Claudia 1975. The Oral Interview Test. 29-44 in R.L. Jones and B. Spolsky (eds.). Testing Language Proficiency. Arlington, Virginia, Center for Applied Linguistics.